Data Management Plan

This Data Management Plan (DMP) describes the elements and procedures for storing, securing and sharing data associated with the Collaborative research described in this proposal and long term preservation and access for data from the C-Mod facility.

1. Data Covered: This plan covers the following data:

- Information from collaborations
  - Any data created as a product of collaborative research, described in this proposal and not stored off-site and covered by that site's DMP
- C-Mod information
  - Continuously acquired engineering data from the C-Mod supervisory control systems
  - Raw data acquired from each C-Mod pulse (shot)
  - Data processed automatically or manually and stored in the MDSplus archive
  - High-level analyzed C-Mod data, run information and the electronic logbook stored in relational databases
  - Modeling and processed data results
- Publications and research reports in digital form

2. Data Acquisition, Storage, Archival and Retention Policy

2.A. Data Acquisition

Data from the C-Mod tokamak has been acquired through the MDSplus system (www.mdsplus.org). This system integrates all set-up, raw and processed data into a single, coherent hierarchical structure. Engineering data is acquired from the supervisory control systems, which is based on programmable logic controllers (PLC) monitoring about 750 distinct system values at a typical acquisition rate of 1 per second on a 24/7 basis. These data are transferred to MDSplus every 2 minutes. Most raw data from experiments is acquired by transient recorders and other devices and written into MDSplus immediately after each pulse. During experiments, pulses occur at the rate of ~4 per hour for 8 hours each run day. Between shots, a large number of automated analysis routines create and store processed data. Together, these amount to 10 GB of data per shot for each of approximately 2,000 shots per year. Over the ensuing days, months and years, additional analyzed data is written into the same data hierarchy. Relational databases are used to store a comparatively small quantity of run metadata, highly analyzed data, data summaries and an electronic lab notebook. Together, integrated over all C-Mod operation, these comprise on the order of 2-3 million database records, most entered manually or by user provided software.

2.B. Data Storage

For critical data, primary storage currently consists of a 110 TB RAID6 disk array that contains all "user" data and all of the C-Mod data taken from its first day of operation. Every night, any new or modified data is copied to a secondary 110 TB disk array, maintained in a separate building.

2.C. Data Backup and Archival

An archive copy of all original C-Mod MDSplus data – that is in the state when it was first acquired – is maintained by nightly transfers to MIT's Tivoli Storage Manager (TSM), a large enterprise-class automated tape library. In addition a backup copy of the C-Mod data and all user files, in their current state, are also maintained and updated daily on TSM. Older versions of files are maintained on TSM for 30 days, allowing further redundancy and a path for recovery from short-term data integrity problems. Additionally, user files are saved monthly for 1 quarter, quarterly for first year, annually after that on TSM. The relational databases are backed up nightly, weekly, and monthly; nightly and weekly backups are saved for 8 weeks and the monthly backups are saved permanently. All of these database backups are in turn backed up to TSM on campus. Individual desktop systems in user offices are backed up using MIT's CrashPlan cloud service.

2.D. Long Term Data Retention Policy

All data ever acquired from C-Mod is stored on magnetic disk and archived as described above, which provides for 4 copies of raw data and 3-4 copies of processed data, spread over 3 separate buildings on campus. While there is no contractual obligation to retain this data beyond the life of the C-Mod cooperative agreement, a recently funded proposal supports data retention and access. Our intention is to keep this data accessible and usable beyond this agreement, limited only be future fiscal constraints, for at least 3-5 years beyond the life of the experiment. Critical user files will be retained on the same basis.

3. Data Access and Sharing

MDSplus provides access to all of the experimental data described above through a simple application program interface (API) adapted for many common programming languages. Data is specified by a globally unique name, decoupling users from details of the underlying storage mechanisms. Remote access is provided by MDSIP, a software-based network layer that allows the API to store or retrieve data using the internet IP protocol. All C-Mod data is available to everyone on the C-Mod team, subject to the use and publication conditions described in the C-Mod collaboration agreement.

http://www.psfc.mit.edu/research/alcator/program/collab_agree_2.pdf

Access to processed data from other machines will be subject to collaboration and data sharing rules specified by those facilities.

4. Publication of Documents and Digital Data

DOE's Office of Science has recently established a new policy for management and access to digital data created through federally funded research. The requirements refer to research products, which include both digital data and digital documents. Of particular note, the policy includes a substantially new requirement for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication". The phrase "data displayed" refers specifically to figures and tables within the publication.

4A. Open Access Data Management

The Harvard/MIT Dataverse data repository will provide a stable, long-term, open, institutional archive for the digital data required under the new rules. To meet the requirement for open-access to data, researchers will create a set of data files that correspond to the figures and tables as they prepare manuscripts,. We have chosen to standardize on the HDF5 file format for data in figures and plain text or Excel for tables. Software for creating these files from within user applications in commonly used scientific programming languages will be provided. Users would submit these files to the PSFC library who will administer the process and organize the data files within the repository.

To provide meaning and context, two general types of metadata will be associated with these data files. The first type provides a description of the data within the files and the second type describes the data collection and associated publication. Inclusion of the first type of metadata will be supported through the software used to write the files. The second type of metadata will be supplied by the authors when the data is submitted through the PSFC web site.

4B. Document Management

To meet differing and evolving requirements from funding agencies and from MIT, digital documents will be stored redundantly in several systems. The PSFC Library will administrate the deposit to DOE P.A.G.E.S. of required metadata and links to full text as specified by OSTI/DOE. (Per the requirements of the *DOE Public Access Plan*). P.A.G.E.S. will link to a full-text version of the accepted manuscript twelve months from the article publication date and then link to the VoR when and if it becomes available. Metadata accompanying the accepted manuscript, e.g., author name, journal title, and digital object identifier (DOI) for the VoR, ensures that attribution to authors, journals, and original publishers will be maintained. All curated document versions are accessible through the PSFC Library website and data repository – these versions are considered "published" once they have been processed and administrated through the PSFC Library document ecosystem, which includes: deposit into the PSFC local digital archive; catalogued (to include all metadata) in the PSFC Online Public Access Catalog (OPAC) which includes links to all document versions; and deposited into MIT's DSpace.

Curated Document Versions can include:

a) Unabridged Manuscript
b) Non-Peer Reviewed Preprint
c) Peer Reviewed Preprint
d) Research or Internal Reports
[both Peer- and non-Peer-Reviewed]
e) Revisions, Errata if any
f) Final version of manuscripts as published
g) DOE regarded Version of Record (VOR), i.e. published version
h) Data sets as shown in final manuscript figures and tables (as discussed in 4A)

To ensure long-term preservation and access, all DOE-funded authors at the PSFC will be required to submit an accepted manuscript and its associated metadata to the PSFC Library for storage on the local (PSFC) domain-specific data archive. Backups of this local archive are done with TSM. Further redundancy is achieved via the PSFC Library depositing the document into MIT's DSpace institutional repository. DSpace is an open source repository used for creating open-access for scholarly and/or published digital content. As a digital archives system, it is focused on the long-term storage, access and preservation of digital content and as such is committed to upholding industry standards of digital curation and preservation principles; it is underwritten by MIT's commitment to provide ample resources to ensure its continued operation. Under the auspices of MIT, the PSFC document archive is assured continuance and/or migration to DSpace through institutional support.

Both the PSFC archive and the MIT DSpace repository are open-access to all, within and beyond the PSFC and MIT communities, without restrictions, aside from any copyright or terms-of-use provisions that may apply to specific documents or data sets.